Flyber Data Strategy MVP

Introduction

Flyber has been massively successful. Results have beaten expectations and projections! This is good news for Flyber, but now it's time to plan for what's next. With success came some challenges. While we were able to grow, the original data pipelines to receive and process data are unable to keep up with the current and future growth.

As a Data Product Manager, working with multiple teams and stakeholders is imperative to success. To understand what our needs are, what scale we are growing at, and how we can build for the future, we need to consider all relevant stakeholders. In this proposal, present your findings along with the analysis and reasoning behind the choices made in order to help Flyber continue its success.

Section 1: Data Customers & Needs

Flyber is a two-sided platform. You have customers who are riders, and you have partners who are drivers/pilots (think Uber: riders and drivers). For the Minimum Viable Product, you will be focusing on the Riders side of the business. To build an end-to-end data pipeline the very first step is to understand who needs data and why they need that data. Within Flyber, identify who your primary data customers/stakeholders are, why they are your primary data stakeholders and how they want to use the data (primary use-cases).

Identify your primary internal stakeholders and their use-cases:
(You may add more rows if necessary.)

Stakeholder	**Why are they primary stakeholders?**	**Use-Case**
Engineering	As a product led business it's important that engineering decisions are led with data. Therefore, we need to include them as a primary stakeholder.	Identifying resourcing needs for android, iOS or web application depending on customer usage. Identifying popular feature usage and un-popular features for prioritizing future development
Marketing	Flyber want to become the number 1 taxi service used in high density areas of NYC, through	Segregate usage by neighborhood

	repeat booking of riders who want to get to their destination faster. Therefore, it's vital that marketing have access to our data, so that we can target certain demographics and areas of NYC to meet our goals	Daily active users by neighborhood Repeat users by neighborhood
Investors	Investors are an important data stakeholder because as a start-up you often need to obtain funding to continue, especially in the flying taxi industry. We will need to demonstrate to our investors that we are achieving our KPIs / Goals	Sharing daily active user data with Investors Sharing popular features with investors Sharing demographic data with investors e.g. neighborhood

Section 2: Data Collection and Data Modelling

To support our primary stakeholders's use-cases we need following data:
(You may add more rows if necessary.)

Stakeholder	**Use-Case**	**Data**	**Why is this the primary use-case?**
Engineering	Identifying resourcing needs for android, iOS or web application depending on customer usage.	Device_type user_uuid	This is one of the primary use cases for Engineering as it allows us to resource our teams correctly, ensuring we develop features that work on the devices that our customers use.
Marketing	Daily active users by neighborhood	User_uuid user_neighborhood event_time	As mentioned, Flyber want to become the number 1 taxi service in high density neighborhoods. It's important that marketing have access to the DAU by neighborhoods to support marketing activities.

Investors	Sharing popular features with investors	User_uuid event_page event_type	A primary use case for Investors will be understanding which features of the product are doing well and are popular. Investors will want to understand whether new product features are working well and deserve to be continued to be funded.

The tables we need are:

Note: As a best practice, we should establish these relationships between tables from the very beginning. To complete this exercise we will focus on fundamental concepts of relational databases - tables, normalization and unique keys. Please provide the table header row for each table, tables might be different lengths. Make sure you include the following for each table. You can create as many tables as you feel are necessary (copy and paste from one of the table sections):

Table 1:

customerData

(You may add more columns if necessary.)

User_uuid	*(There is no foreign key)*	*First Name*	*Last Name*	*Neighborhood*

Rationale for Choosing Primary and Foreign Keys for the Table 1:

I feel that it's important to start with the customerData table. The table would normally hold a lot more information but I don't have that available.

- The reason for choosing the user_uuid as the primary key is that it's a unique identifier for a customer.
- I did not add a foreign key as I don't believe other than the user_uuid there would be another unique field used a primary key in another table. Therefore I left it out.

Table 2:

Ride History

(You may add more columns if necessary.)

Ride_uuid	*User_uuid*	*user_neighborhood*	*user_neighborhood_dropoff*	*Ride_Duration*	*Device_type*

Rationale for Choosing Primary and Foreign Keys for the Table 2:

- The primary key here was ride_uuid as each ride is unique
- Foreign key was user_uuid as it's the primary key from customerData table and each unique ride_uuid will have a user_uuid associated to it.

Table 3:

Channel Analytics

(You may add more columns if necessary.)

Event_uuid	*User_uuid*	*Device_type*	*Session_Uuid*	*Event_page*	*Event_type*

Rationale for Choosing Primary and Foreign Keys for the Table 3:

- The primary key for this table is the event_uuid as it is unique.
- The foreign key for this table is again the user_uuid as it's a primary key of the customerData table and each unique event_uuid relates to a user_uuid.

Section 3: Extraction and Transformation

Now that you have the requirements from your stakeholders, you want to understand the current state of what data is collected. That is how you recognize which additional data you need to achieve the future state. You ask the engineering team what data they are currently collecting in the pipelines, and they provide you with section_3_event_logs template (which you can download from the classroom) generated by rider's activities on the Flyber App. Also provided in the Project Resources.

Extraction and Transformation-1

ETL is performed on the provided Event Logs Template and results will be transferred to the proposal template. The project's ETL should be created inside of your copy of the Event Logs template in the tab titled, ETL. Clicking on the link above will create a copy of the Event Logs for you

After being provided with a CSV log file, use extraction techniques to be able to get the data into a usable form. Because this needs to be a repeatable process we need to document it in order to assess its feasibility. Below,

1. Write the steps you took to extract the data and provide reasoning for why you used this method *Note: Don't forget to include any file type changes*:
2. Perform cleaning and transformation of the data in the ETL tab and document.
3. Document and provide rationale for all your steps below as well.

Steps for Extraction:

- Extracted the data into separate pivot tables to allow for analysis to occur.
    - The data was structured enough that pivot table analysis was possible and the most efficient manner at this time.
- Extracted the daily data for visitors into separate tabs to allow for DAU calculations to be completed in Tableau later.
    - Due to limitations of distinct count on Mac I will have to use Tableau to analyze the daily active users.

Transformation-2

Analyze the data from part 1 to answer the following questions:

1. How many events are being recorded per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Event Count	9891	18056	18202	17963	17600	17694	17595

2. How many events of each event type per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Choose Car	1498	2843	2953	2769	2725	2801	2804
Search	1484	2891	2824	2899	2749	2904	2821
Open	6594	11733	11767	11662	11531	11325	11371
Begin Ride	38	49	62	86	57	57	78
Request Car	277	540	596	547	538	607	521

3. How many events per device type per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
ios	2384	4337	4217	4373	4380	4482	4500
android	1463	2870	2854	2729	2744	2562	2672
Desktop Web	895	2007	1600	1958	1712	1866	682
Mobile Web	5149	8842	9531	8903	8764	8784	8646

4. How many events per page type per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Search Page	3995	7219	7307	7221	6979	7201	7137
Book Page	1977	3548	3576	3572	3586	3424	3506
Driver Page	1977	3548	3576	3572	3586	3424	3506
Splash Page	2954	5466	5448	5376	5280	5380	5184

5. How many events for each location per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Manhattan	6869	12591	12807	12180	12270	12371	12201
Brooklyn	2009	3737	3590	4025	3440	3400	3556
Bronx	250	533	507	469	510	394	558
Queens	595	842	905	893	1026	1069	936
Staten Island	168	353	393	396	354	460	344

ETL Automation and Scalability:

Provide an analysis about this ETL process. Address and provide rationale for manually extracting, loading, and transforming the data from the raw logs. Also address potential preliminary recommendations on improving this process.

- It felt like a good manual approach using excel and mainly using the pivot table functions. I was able to get some good insight into the data with minimal effort.
- The data was relatively clean and I did not perform any cleaning functions on it.

- The data was structured making it easier to ETL manually with excel.
- For what I was asked to do with the data it made sense to ETL this manually as it was not required to be analyzed in real time.
- An improvement to the process would be the ability to use distinct count when using pivot table analysis. Due to the large numbers of repeat user_uuid and session_uuid and the inability to use distinct count on Mac I was unable to easily perform some actions that I wanted to.

Section 4: Choosing Relevant Dataset

The previous exercise gave you a sneak peek into the Extraction and Loading aspects of ETLs in data pipelines. For making business decisions, a data consumer would like to have all the data they want. However, for any ecosystem, it is impossible to collect or provide everything that the customers need. In this exercise, you will get a taste of real-world scenarios wherein:

- All the resources are not always available to get what you need.
- You have to get creative and get the most insights with a minimal data set.

Oftentimes your stakeholders/customers will "ask for the moon", but you'll have to push them to work with the small amount of information you have and get creative.

Note: As you learned in the course, being a Data Project Manager involves an extraordinary amount of collaboration. Complete the next sections based on the following scenario.

After the analysis in section 3, we made sense of the numbers, and realized the total number of events seems to be too small (this was a week's worth of data, but you need at least a month). Further investigation reveals that this was a subset of logs, but the actual data that is being collected is much bigger. Working through this small data set was tedious and repeating this exercise on a much bigger data set manually won't be feasible. Considering the time constraints of this project, engineering is willing to help with some automation. They also have limited bandwidth and are busy scaling systems up.

Engineering is willing to provide some data, but they have asked for the criterion that is most important. To First provide your business question and provide a rationale for why this is the most important.

Business Question: Which locations do we need to increase our marketing activities to attract more bookings?

Choose one of the following prompts that you think can get you the most relevant information to proceed further.

1. How many events are being recorded per day?
2. How many events of each event type per day?
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

For your chosen question also answer the following using the data from section 3 to support your answer:

1. How much is the customer data increasing?
2. How much is the transactional data increasing?
3. How much is the event log data increasing?

Which of the following data is ***most*** important to answer this question? Why?

- Event Log Data
- Transactional Data
- Customer Data

To answer our question, we need to see event log data for each location so that we can analyse the amount of requests and booking that are fired at each location. Combined with the event page data, we will be able to create insights that allow us to understand things like:

- Ride requests per location, per day
- % of visits to booking page, that result in a ride request, per location
- Repeat ride requests per location

All of this data will help our marketing team position their marketing efforts in the right way to help Flyber become the number 1 taxi service in high density neighborhoods of NYC.

Section 5: [Optional] Loading and Visualization On Your Own

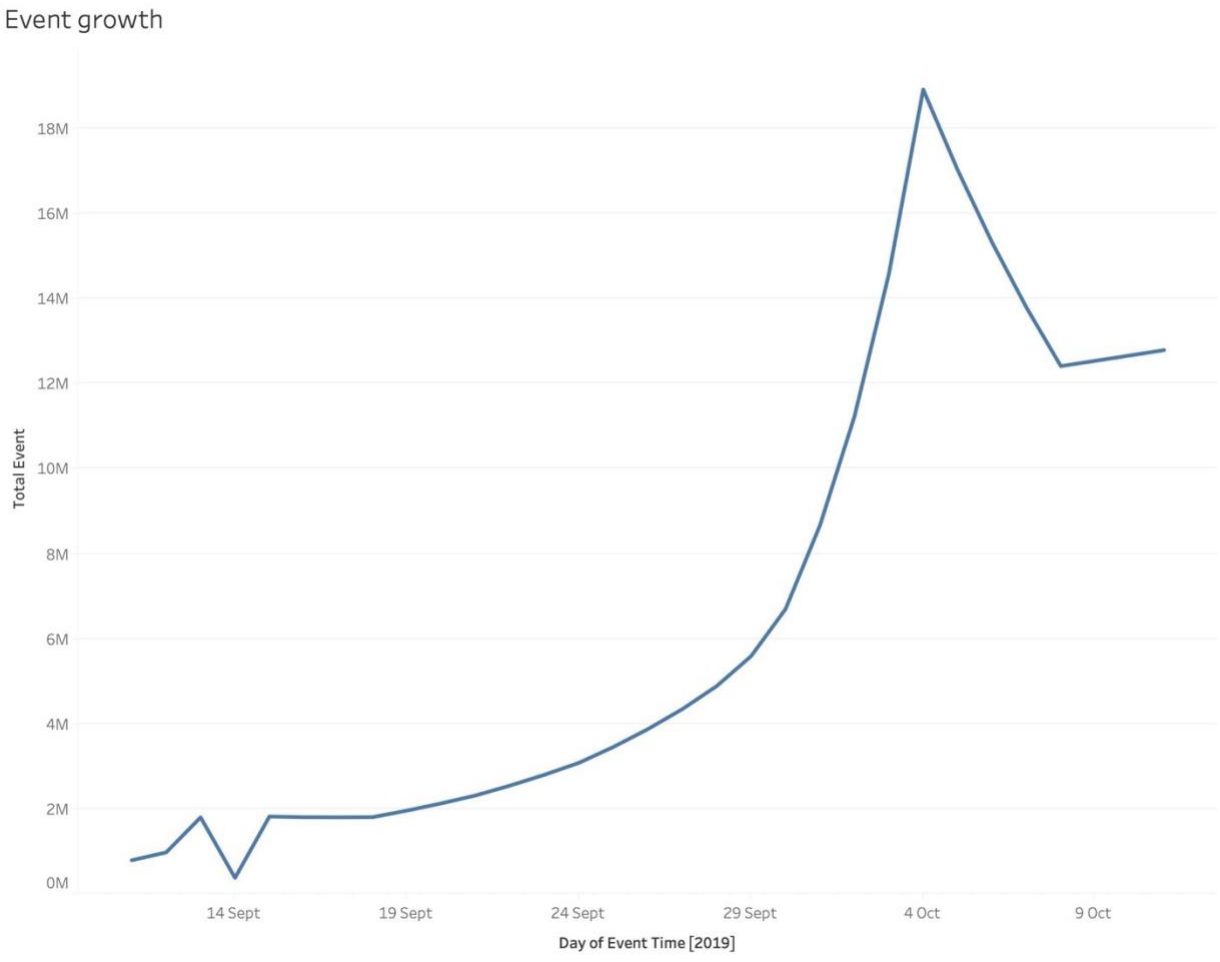
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

After sharing your criterion with engineering, they give you a new set of data: Section 5 Event Type Log also available in the classroom resources. Also provided in the project resources section.

Engineering provided you with the data you want, but you have yet to achieve your goal as a Data Product Manager. Now, utilize the data to make business decisions. Your executives do not want you to give them a bunch of data tables; instead, they prefer visualizations to help convey the key insights succinctly. Visualizing this data will help you understand the underlying trends and help you determine the story that needs to be told in your proposal to executives.

In this section, you can load and visualize the data into whatever platform you would like. A Python Notebook, Tableau, or any other visualization tool you are familiar with. Create two visualizations that might help you to better understand your data trends and place either a screenshot or exported image of your visualizations and the details of each below. Please provide the steps you took to visualize your data and what the visualization tells you about your data.

Visualization 1:



Event growth over time: This graph tells us that between 14th September and 4th October there was a large and consistent growth in the total amount of events generated. Then from 4th October to 9th October there was a sharp decrease in the events generated. When more events are generated that equates to higher traffic through our channels and vice versa when events decline.

This graph was created using the following steps:

1. *Adding event time as a column and changing it to display the days*
2. *Adding the sum of the total events along the rows*
3. *Changing to a line graph*

Visualization 2:

Difference in total from previous day during sharp decline of events

	Day of Event Time				
	4 October 2019	5 October 2019	6 October 2019	7 October 2019	8 October 2019
Difference in Open from t..		-1,248,376	-1,123,538	-1,011,184	-910,065
Difference in Search from ..		-309,916	-278,925	-251,032	-225,929

Difference from the previous day's events focused on open and search: This graph tells us the difference from the previous day's events total for certain event types. I focused on open and search as from the events that we have, those are the events that are generated at the beginning of the customer's journey using our business.

Difference in total from previous day during sharp decline of events

	Day of Event Time				
	4 October..	5 October..	6 October..	7 October..	8 October..
% Difference in Open fro..		-10.00%	-10.00%	-10.00%	-10.00%
% Difference in Search fro..		-10.00%	-10.00%	-10.00%	-10.00%
% Difference in Request C..		-10.00%	-10.00%	-10.00%	-10.00%

What is clear from the data is that the drop-off is consistent across the important event types. You can see when comparing the % difference that all events are running at a 10% difference from the day previous. This tells us that the drop-off comes at the top of the funnel (e.g. getting users to open the app) rather than a problem with requesting a car. If open and search events remained the same as the previous day but request a car dropped 10% we could conclude that there might be an issue with the request a car feature in our app.

This graph was created using the following steps:

1. *Adding the days as the columns and filtering on the date period that had the sharp decline in events generated.*
2. *Including certain events that we wanted to analyze. Measuring their value and creating a table calculation that showed the difference and % difference.*

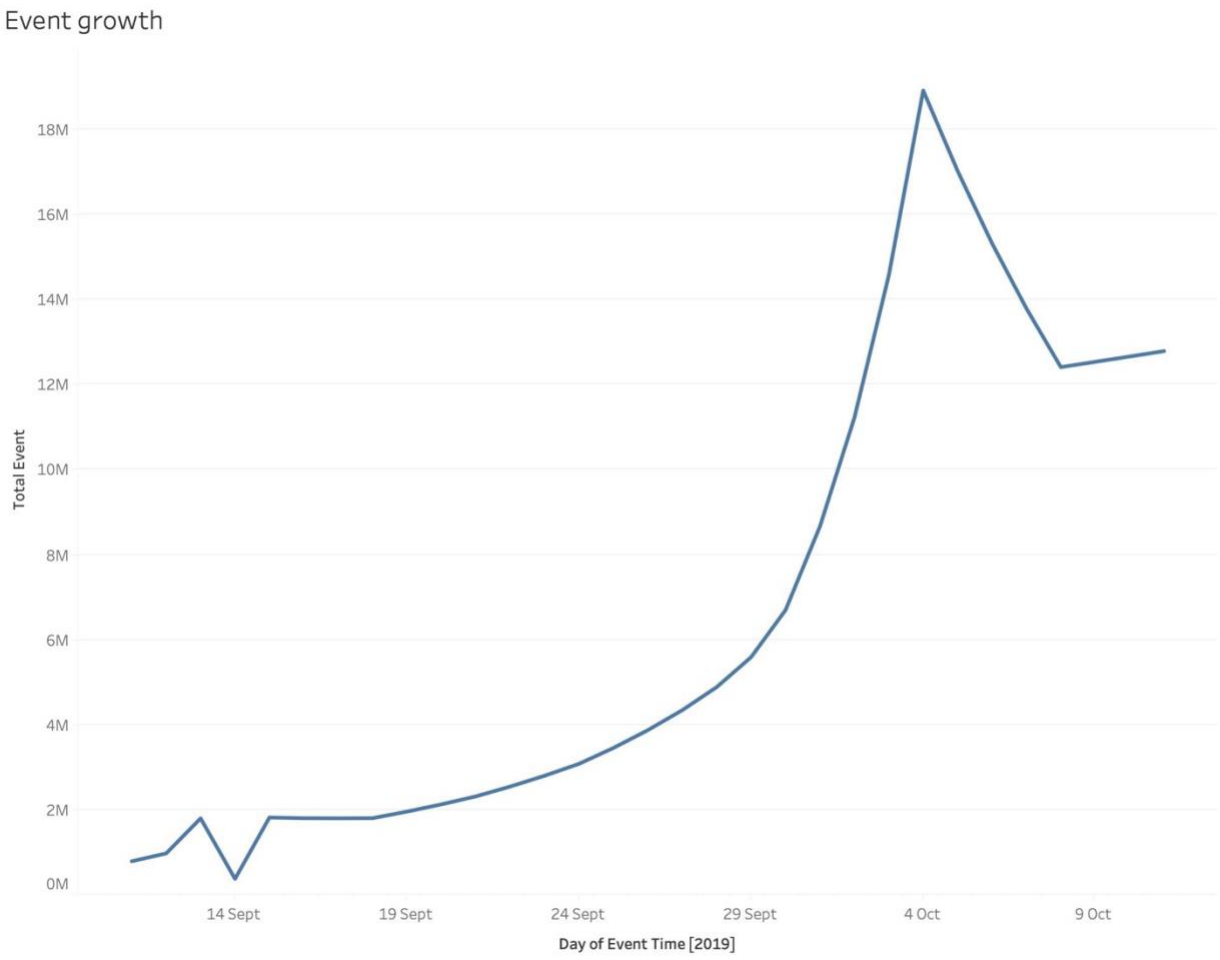
Section 6: Business Insights

The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received and considered more robust.

What is the story the data is telling you about Flyber's data growth? If you created Visualizations, you can use them as well, but they are not required). Include any data and calculations that were made to help tell that story and quantify the data growth.

Data Growth for Last Month

Visualization:



Data and calculations used for quantifying of Flyber's Data Growth:

Event growth over time: This graph tells us that between 14th September and 4th October there was a large and consistent growth in the total amount of events generated. Then from 4th October to 9th October there was a sharp decrease in the events generated. When more events are generated that equates to higher traffic through our channels and vice versa when events decline.

This graph was created using the following steps:

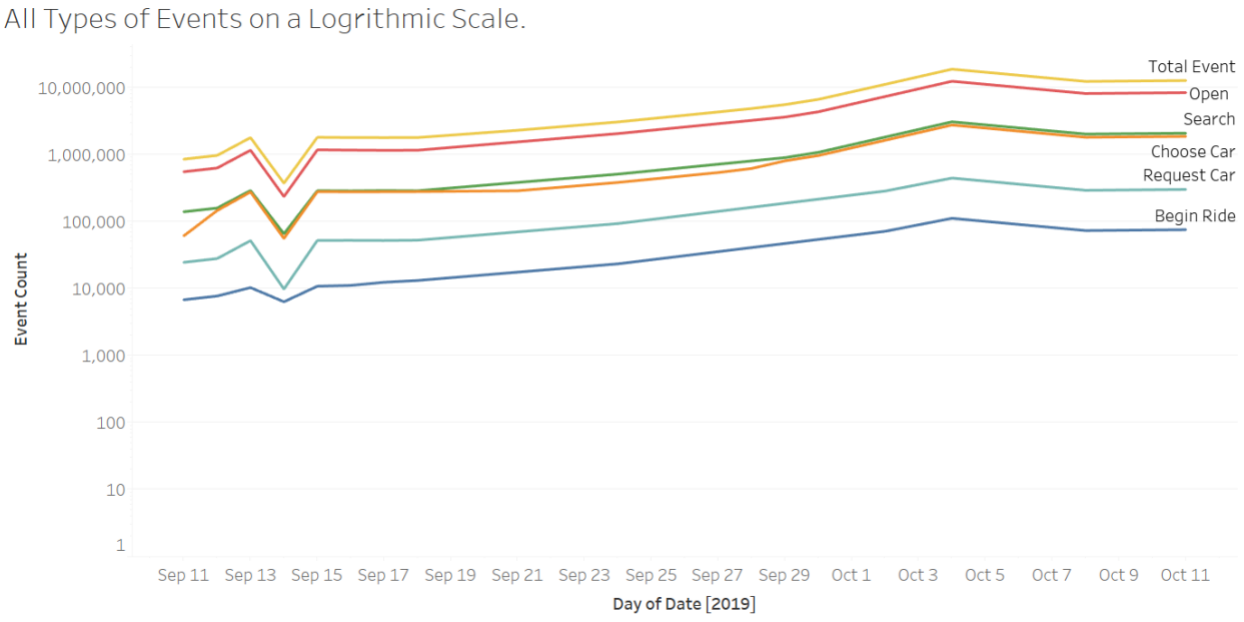
1. *Adding event time as a column and changing it to display the days*
2. *Adding the sum of the total events along the rows*
3. *Changing to a line graph*

What is the fastest growing data and why?

We cannot assert which is the fastest growing as we just have access to event logs and not transactional data or customer data.

All Event Type Data

Visualization:

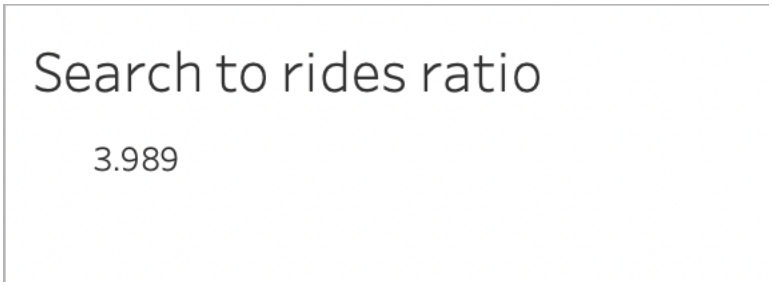


What is the Data Story our data tells for each of the following:

- Graph Pattern
- Good or Bad
- October Marketing Campaign
- Marketing Campaign Impact
- Importance of Relationship Between Marketing Campaigns and Data Generation

- *The graph pattern is consistent across events from left to right. To me that signifies that our customer journey is quite consistent and that our features work once we get customers into the funnel. The fact that Open has a similar graph pattern to begin ride, suggests that our Ride to Open ratio are related to each other. If our open pattern went higher to the right than our begin ride pattern, we could argue that that customers were dropping out of the funnel before beginning rides*
- *Based on my analysis in the previous bullet point I consider this graph pattern to be good; there are certain patterns on the graph that are promising as well. For example, when all previous events generated in the funnel before begin ride take a sharp decline at the beginning, the begin rides events do not decline as sharply, indicating that rides continue and aren't as directly impacted by the funnel.*
- *October's marketing campaign was not effective. We don't have data to know whether there was a marketing campaign during September but the marketing campaign during October resulted in a decline in growth and a decline in rides. There are many other factors that could impact this and it would require additional data to be sure (e.g. seasonality)*
- *The impact that our marketing campaign had was a lower number of events generated across the data points that we have at our disposal. 1 question might be whether the marketing campaign is directing customers to other parts of the application e.g. customer sign up events. That is unclear to us but what we can say is that it had a negative impact on rides.*
- *Marketing campaigns cost money and they should always have a goal. For example we could run a retention marketing campaign whereby we expect to increase the number of repeat bookings. For us to understand the success of that marketing campaign we have to hypothesis on the outcome before and to then confirm or deny that hypothesis we need access to the events generated. In this example having access to user_uuid, begin_ride and other events during a marketing campaign would allow the business to decide whether the campaign is having a positive effect.*

Search to Ride Ratio



For approx. every 4 search events generated we generate 1 begin ride event. This could give us a search to ride ratio of just under 4 searches for every ride booked.

Section 7: Data Infrastructure Strategy

Thus far we have:

- identified data stakeholders and their data needs.
- Identified what data is currently being collected and what data needs to be collected.
- Identified data insights and growth trends.

Now, it's time to tie all the loose threads together and bring this process to its logical conclusion by suggesting which Data Warehouse (DWH) Flyber should invest in and why. Using data warehouse options below, suggest whether Flyber should choose an on-premises or Cloud data warehouse system and which specific data warehouse would best serve Flyer's data needs.

Data Warehouse Options:

Cloud:
- Amazon Redshift
- Google BigQuery
- Snowflake
- Microsoft Azure

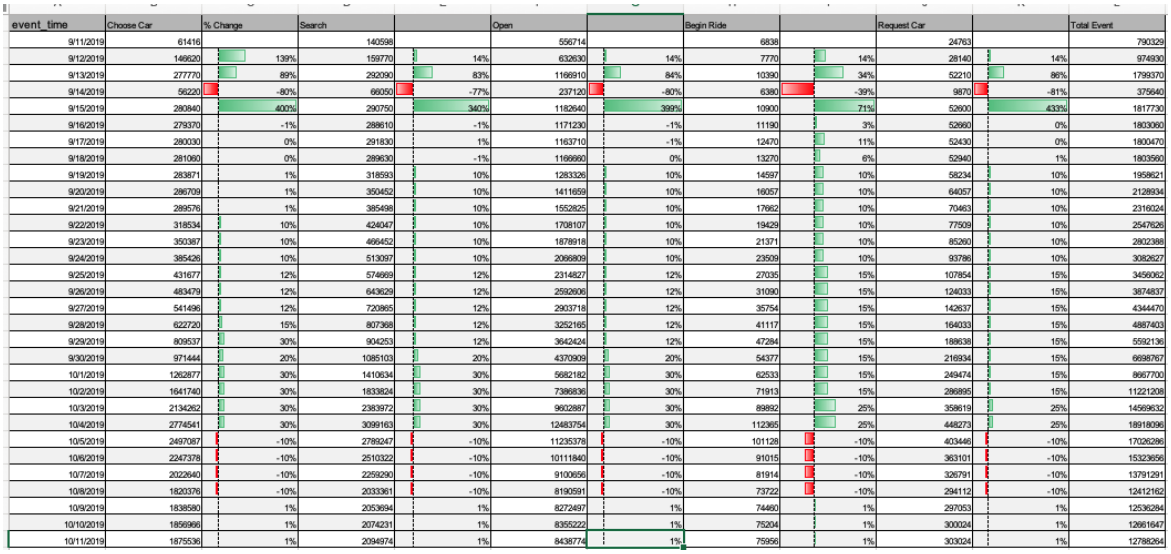
On-Premise:
- Oracle Exadata
- Teradata, Vertica
- Apache
- Hadoop

You will address the following factors with a rationale as to why the DWH chosen is the best for Flyber:
- Cost
- Scalability
- In-house Expertise
- Latency/Connectivity
- Reliability

Cloud vs On-Premise
Provide an evidence based solution as to why Flyber would be best served by a Cloud or on-premise DWH. In this response, you don't need to specify *which* specific Cloud or on-premise DWH product you will choose, just if it will be Cloud or on-premise. Remember to address the factors above.

I would advise that Flyber invest in a cloud based DWH for the following reasons:

- Scalability. We have seen the large increase in events generated and sometimes also the decrease in the events created over a 1-month period. Focusing on the increase, the largest singular increase for total events generated from the previous day for an event type was 433%. The advantage of cloud vs on-premises in this instance is that scale can be handled automatically and increases like this will be auto-scaled to not impact performance.



event_time	Choose Car	% Change	Search		Open		Begin Ride		Request Car		Total Event
9/11/2019	61416		140598		556714		6838		24763		790329
9/12/2019	146620	139%	159770	14%	632630	14%	7770	14%	28140	14%	974930
9/13/2019	277770	89%	292090	83%	1166910	84%	10390	34%	52210	86%	1799370
9/14/2019	56220	-80%	66050	-77%	237120	-80%	6380	-39%	9870	-81%	375640
9/15/2019	280840	400%	290750	340%	1182640	399%	10900	71%	52600	433%	1817730
9/16/2019	279370	-1%	288610	-1%	1171230	-1%	11190	3%	52660	0%	1803060
9/17/2019	280030	0%	291830	1%	1163710	-1%	12470	11%	52430	0%	1800470
9/18/2019	281060	0%	289630	-1%	1166660	0%	13270	6%	52940	1%	1803560
9/19/2019	283871	1%	318593	10%	1283326	10%	14597	10%	58234	10%	1958621
9/20/2019	286709	1%	350452	10%	1411659	10%	16057	10%	64057	10%	2128934
9/21/2019	289576	1%	385498	10%	1552825	10%	17662	10%	70463	10%	2316024
9/22/2019	318534	10%	424047	10%	1708107	10%	19429	10%	77509	10%	2547626
9/23/2019	350387	10%	466452	10%	1878918	10%	21371	10%	85260	10%	2802388
9/24/2019	385426	10%	513097	10%	2066809	10%	23509	10%	93786	10%	3082627
9/25/2019	431677	12%	574669	12%	2314827	12%	27035	15%	107854	15%	3456062
9/26/2019	483479	12%	643629	12%	2592606	12%	31090	15%	124033	15%	3874837
9/27/2019	541496	12%	720865	12%	2903718	12%	35754	15%	142637	15%	4344470
9/28/2019	622720	15%	807368	12%	3252165	12%	41117	15%	164033	15%	4887403
9/29/2019	809537	30%	904253	12%	3642424	12%	47284	15%	188638	15%	5592136
9/30/2019	971444	20%	1085103	20%	4370909	20%	54377	15%	216934	15%	6698767
10/1/2019	1262877	30%	1410634	30%	5682182	30%	62533	15%	249474	15%	8667700
10/2/2019	1641740	30%	1833824	30%	7386836	30%	71913	15%	286895	15%	11221208
10/3/2019	2134262	30%	2383972	30%	9602887	30%	89892	25%	358619	25%	14569632
10/4/2019	2774541	30%	3099163	30%	12483754	30%	112365	25%	448273	25%	18918096
10/5/2019	2497087	-10%	2789247	-10%	11235378	-10%	101128	-10%	403446	-10%	17026286
10/6/2019	2247378	-10%	2510322	-10%	10111840	-10%	91015	-10%	363101	-10%	15323656
10/7/2019	2022640	-10%	2259290	-10%	9100656	-10%	81914	-10%	326791	-10%	13791291
10/8/2019	1820376	-10%	2033361	-10%	8190591	-10%	73722	-10%	294112	-10%	12412162
10/9/2019	1838580	1%	2053694	1%	8272497	1%	74460	1%	297053	1%	12536284
10/10/2019	1856966	1%	2074231	1%	8355222	1%	75204	1%	300024	1%	12661647
10/11/2019	1875536	1%	2094974	1%	8438774	1%	75956	1%	303024	1%	12788264

- This has an important impact on cost as well. Auto-scaling can lead to higher costs as you are outsourcing the scaling to a managed service and an unwanted increase in your data (or a mistake in the code) could lead to a surge in pricing. However, with Amazon RedShift for example you can utilize on-demand pricing that allows you to pay for provisioned capacity by the hour with no commitments. There is also the opportunity to utilize the managed storage pricing whereby you pay for data in a managed storage approach at a fixed GB-month rate. Amazon Redshift Serverless makes it easy for you to run petabyte-scale analytics in seconds to get rapid insights without having to configure and manage your data warehouse clusters. Amazon Redshift Serverless automatically provisions and scales the data warehouse capacity to deliver high performance for demanding and unpredictable workloads, and you pay only for the resources you use.
- Amazon Redshift also replaces the need to deliver our in-house capabilities and could help Flyber move faster in the beginning. They have lots of learning materials that would enable us to utilize our existing data resources and spin-up our DWH in a faster time that if we were to build it in-house.
- Amazon Redshift is fully managed, scalable cloud DWH that can handle structured and semi-structured data.
- **Reliability:** With the Concurrency Scaling feature, you can support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance. When concurrency scaling is enabled, Amazon Redshift automatically adds cluster capacity when your cluster experiences increase in query queueing.
- An on-premises data warehouse provides total control — and total responsibility. Database administrators and analysts, systems administrators, systems engineers, network engineers, and security specialists must design, procure, and install on-premises systems. They must handle moves, adds, and changes — all administration and maintenance of hardware and software. They have full responsibility to ensure that the underlying infrastructure stays up and running efficiently, reliably, and securely.
- Additionally, an on-premises data warehouse cannot accommodate bursts of activity that require more compute or memory. An organization must purchase "up," sizing its data warehouse to handle peak load, even if that level of usage occurs only intermittently. And scaling up to meet changing needs may require replacing systems that cannot meet new demands
- **Security:** Cloud providers invest heavily in physical; and logical security controls. Not only is that a cost advantage, as the costs are distributed across all their customers and built into the pricing models it's also a confidence builder in ensuring our data will be safe.

- **Connectivity / Latency:** Data latency might be an issue but it's one that we can test with our cloud providers. Based on our requirements on performance we might be OK. We don't yet have requirements that need real-time analysis and from the data that we have access to so-far, the analysis is often fine when completed historically. Several factors contribute to latency, such as the location of data sources, quantity of data, and type of data. The best way to assess the impact of latency is to do testing in as close to a production environment as possible.

Suggested DWH

Provide an evidence based solution as to which DWH product is best for Flyber. Remember to address the factors above.

Amazon Redshift

- Amazon Redshift Serverless makes it easy for you to run petabyte-scale analytics in seconds to get rapid insights without having to configure and manage your data warehouse clusters. Amazon Redshift Serverless automatically provisions and scales the data warehouse capacity to deliver high performance for demanding and unpredictable workloads, and you pay only for the resources you use.
- Amazon Redshift also replaces the need to deliver our in-house capabilities and could help Flyber move faster in the beginning. They have lots of learning materials that would enable us to utilize our existing data resources and spin-up our DWH in a faster time that if we were to build it in-house.
- Amazon Redshift is fully managed, scalable cloud DWH that can handle structured and semi-structured data.
- **Reliability:** With the Concurrency Scaling feature, you can support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance. When concurrency scaling is enabled, Amazon Redshift automatically adds cluster capacity when your cluster experiences increase in query queueing.

Image Appendix

Image 1: Log Growth

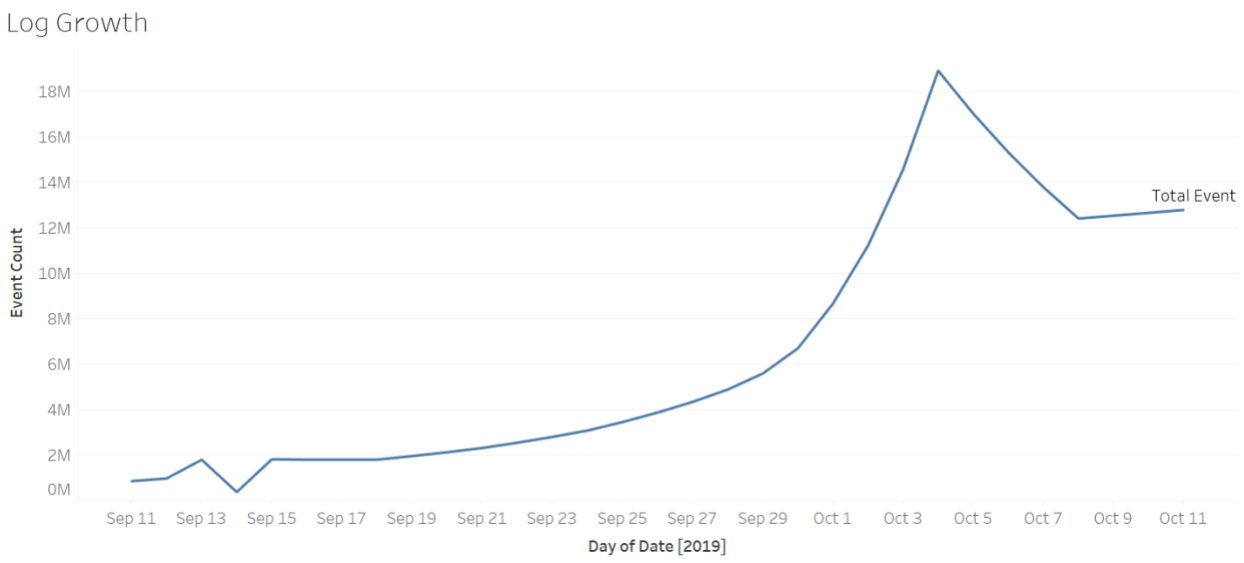


Image 2: Ride Growth

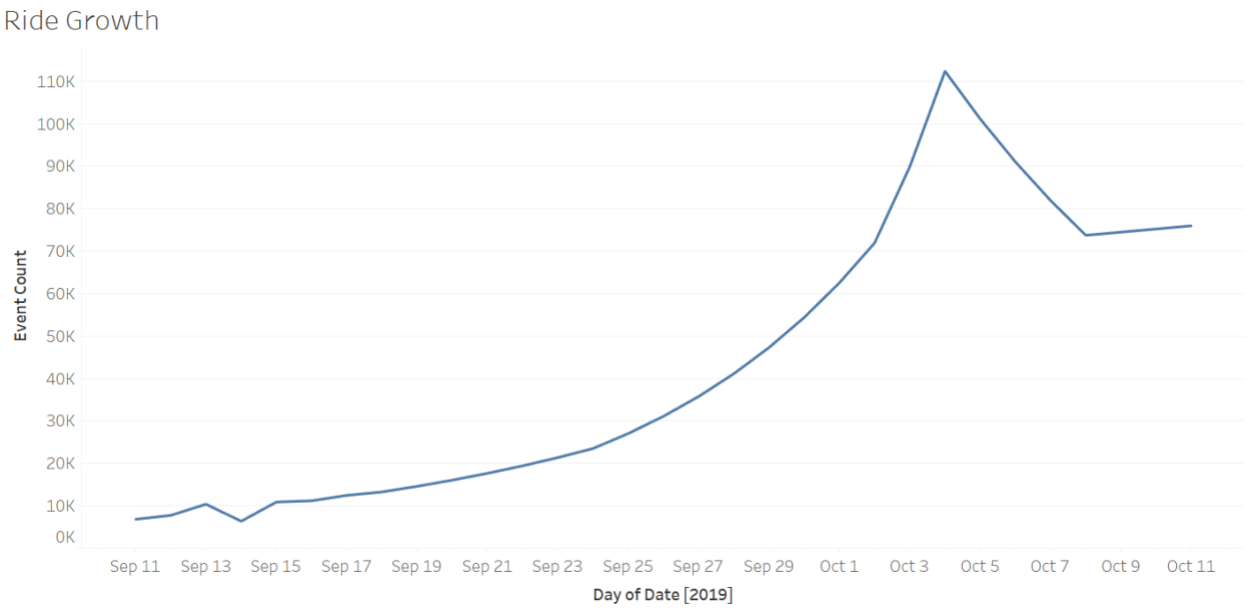


Image 3: Total Event Count

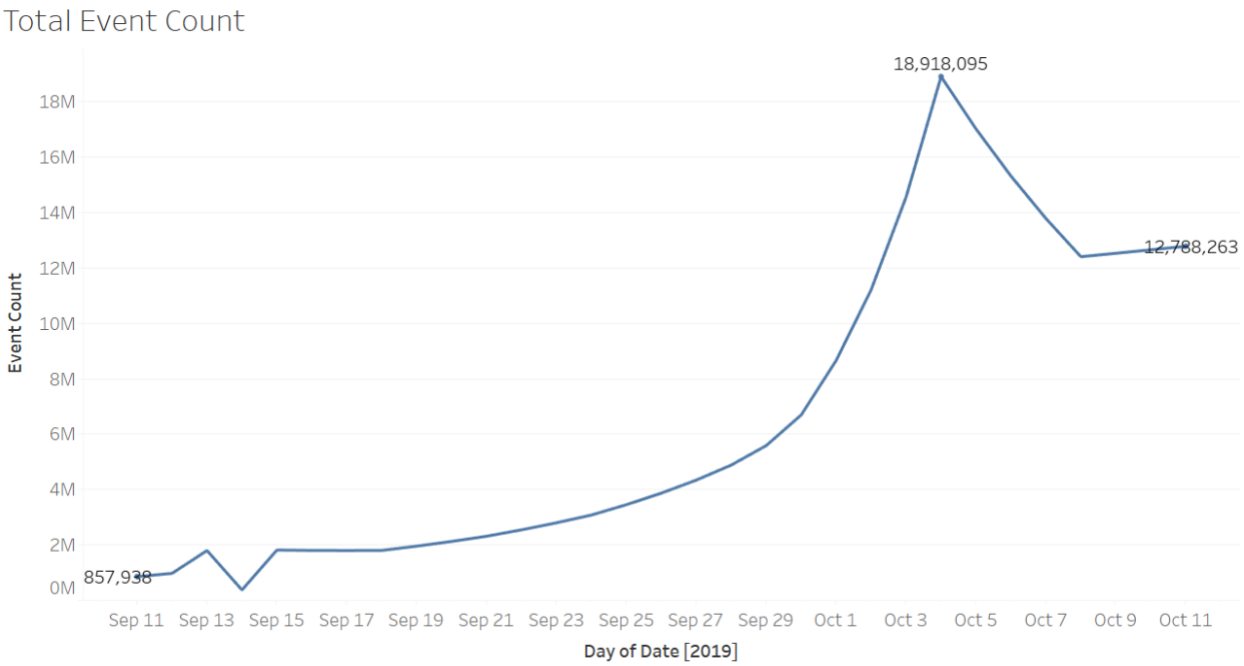


Image 4: All Events Log Scale

All Types of Events on a Logrithmic Scale.